Report of MA678 Final Project

Lintong Li

2022-12-09

Abstract

IMDB is currently an online database of movies, TV shows, etc., which includes names of directors and actors, number of Facebook likes, genres, ratings, vote counts, etc. A commercially successful movie not only entertains the fans but also helps the movie company to earn a lot of revenue. Thus, leaving the question, what factors are so critical in influencing a movie's global revenue? To address this question, I attempted to build a multilevel model with level: genre. The dataset provides a total of 22 raw features, including the movie's budget and revenue. Most of these movies were released between 1999 and 2019. This report contains four main parts: Introduction, Method, Result and Discussion.

Introduction

As IMDB has become the most popular movie rating platform, the amount of data stored in it is very large with approximately 5,000 movies, each with its characteristics, such as genre, duration, budget, global revenue, etc.

Specifically, some famous movies are translated into different languages, such as "Lucy", which was released in 2014 and was converted into 5 other languages: Mandarin, French, Italian, German, Spanish and Korean, while the movies that make the most money are generally the ones with sizable budgets, such as ***action***, ***adventure*** and ***drama*** released in 2019: "The Avengers," with a top budget of 356 million. Due to the large amount of money required for its production and promotion, its impressive presentation can attract more people to watch it and leave positive reviews. Compared to other movies, it has the highest global revenue of 2.8 billion. However, there is a difference in the impact of each genre of movie on revenue.

Therefore, I decided to introduce a multilevel model to explore which factors and how they affect the global revenue of different genre of movies.

Method

Data processing

I found the dataset published on **Kaggle**: Netflix Movies and TV Shows Dataset, named **IMDB movies.csv**

Firstly, I removed the whitespaces of all columns. Secondly, I filtered out the data between 1999 and 2019, probably because movies were not prevalent before 1999, and after 2019, due to COVID, there were fewer opportunities for crowds to gather, leading the movie industry gradually went downhill. Thirdly, I split one row of the genre columns belonging to the same movie into multiple rows, keeping the genre and movie title in one-to-one correspondence. Besides, the budget column was converted into dollar units, and the dollar signs were removed

and converted into numeric types to facilitate data analysis. Then, I log-transformed all numeric columns to make the plot easier to read. Finally, I removed all duplicate values and null values.

Here are some explanations of columns:

column names	explanation
title	ID name of movie
genre	genre of movie
duration	How long is the movie in minutes
avg_vote	IMDB User Rating
votes	Number of votes for the rating
usa_gross_income	USA revenue in dollars
worldwide_gross_income	Global revenue in dollars
metascore	Weighted average of reviews
reviews_from_users	Number of reviews
reviews_from_critics	Number of critical reviews
budget	Amount of money for movie production

Exploratory Data Analysis

I got cleaned data with 12302 observations and 11 variables by processing data, taking global revenue as the dependent variable and 11 independent variables. However, whether every variable is useful or not depends entirely on the following exploring data analysis process.

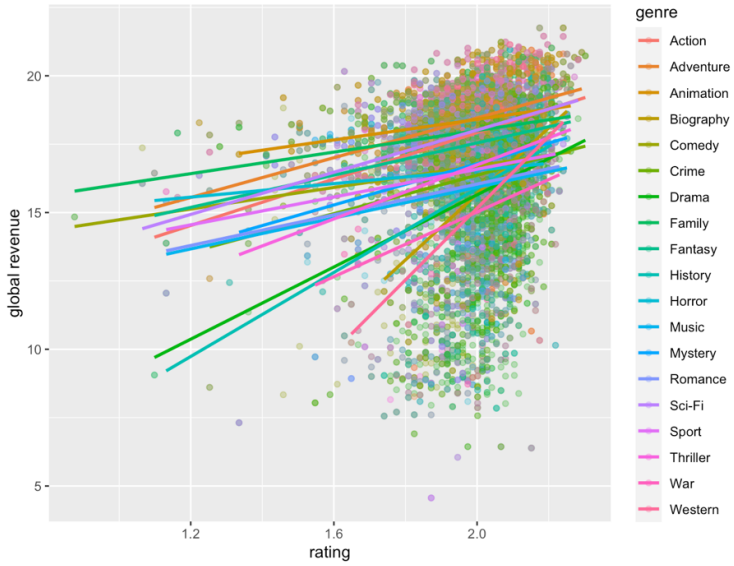

Figure1: relationship between rating and global revenue

Figure 1 shows the relationship between ***rating*** and ***global revenue*** in genre level. Overall, there is an upward trend in the straight line indicating a positive correlation between ***rating*** and ***global revenue***. Looking at each genre, there are differences in the intercept and slope, with the largest slope and smallest intercept for Western films, whose revenues are strongly influenced by ratings.

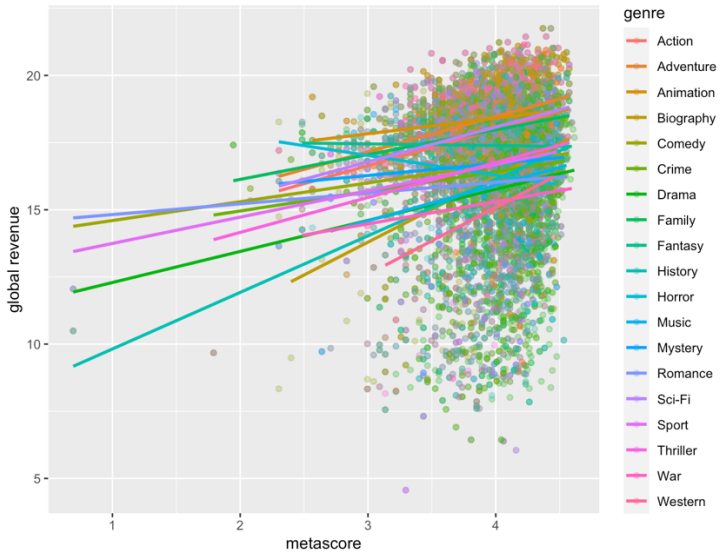

Figure2: relationship between metascore and global revenue

Figure 2 shows the relationship between ***metascore*** and ***global revenue*** at genre level, showing an upward trend. However, the trends in the distribution of meta-scores and ratings are similar across genres, so I believe that these two variables are extremely correlated. Besides, after plotting the relationship between ***global revenue*** and ***duration, votes, budget***, and ***usa revenue***, I found that the trends were almost identical, so I put them in the appendix.

model fitting

Focusing on multiple movie genres, I decided to use a multilevel model to fit **IMDB data**. Unfortunately, continuous data arising do not follow the bell curve, so I applied the log(variable +1) transformation to make it as "normal" as possible. Here is the correlation matrix of all the numerical columns used to determine which predictor variables are selected:

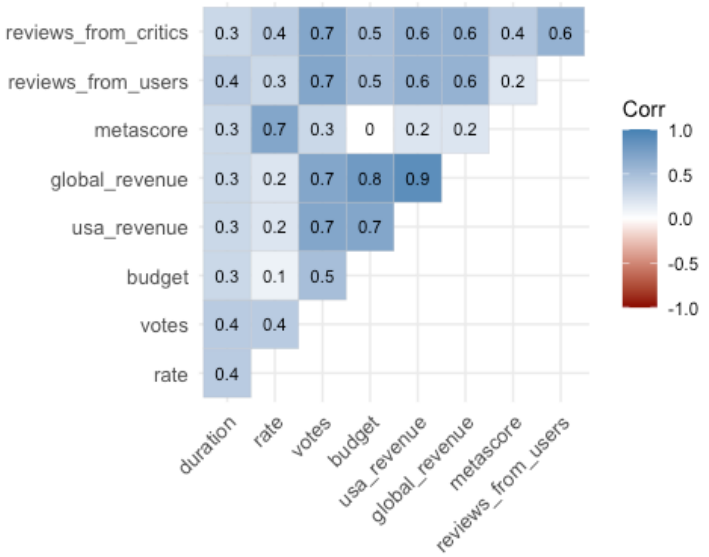

Figure3: correlation matrix

Figure 3 shows the correlation matrix of all potential variables related to global revenue. As an aboved statement, the correlation coefficient between ***metascore*** and ***rating*** is 0,7, which is indeed highly correlated, so I decided to keep ***metascore*** and remove ***rating. usa revenue, votes*** and ***budget*** are the variables that I feel significantly affect the outcome variables, so I kept them directly. However, *reviews from users* and *reviews from critics* are also highly correlated with *votes*, so I removed them as well.

Furthermore, since there is different intercept and different slope in each genre by exploring data, I fitted the multilevel level with a random slope and random intercept, which allows the impact of all predictors to vary randomly from one genre to another. Here is the model I created:

```
model <- lmer(global_revenue ~ duration + votes + budget + metascore + usa_rev
enue +(1 + duration + votes + budget + metascore + usa_revenue| genre), data =
 movieClean)
```

Here is the table of fixed effects, indicating that all variables are considered statistically significant at the $\alpha = 0.5$ level ($p < 0.05$). I also draw a plot to clarify as follows.

Fixed effects:					
	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-2.64	0.44	42.09	-5.98	4.21e-07 ***
duration	0.43	0.10	30.82	4.27	0.000173 ***
votes	0.35	0.02	14.37	16.67	8.46e-11 ***
budget	0.27	0.03	62.23	8.06	3.06e-11 ***
metascore	0.14	0.05	17.90	3.03	0.007272 **
usa_revenue	0.55	0.02	17.71	24.66	3.70e-15 ***

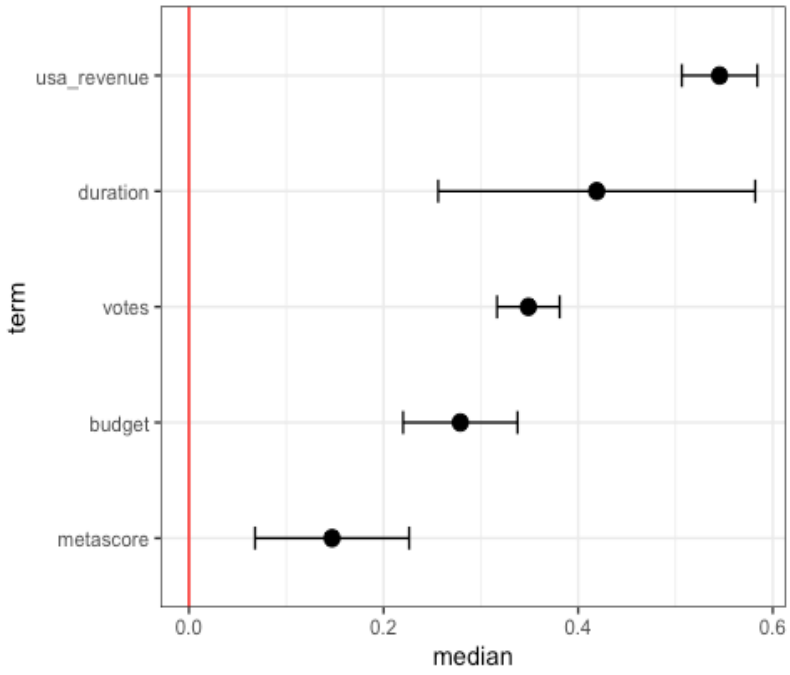

Figure4: Fixed Effects of Model

Here is the table of random effects at genre level and plot the results of a simulation of the random effects to clarify

```
##           (Intercept) duration votes budget metascore usa_revenue
## Action          -0.41    -0.30 -0.06   0.23      0.24       -0.14
## Adventure        0.85    -0.46 -0.10   0.23      0.11       -0.10
## Animation        1.95    -0.35 -0.11   0.17     -0.13       -0.08
## Biography        0.94    -0.12  0.08  -0.01     -0.05       -0.06
## Comedy          -1.32     0.00 -0.11   0.04      0.17        0.07
## Crime           -0.01    -0.13  0.05   0.02      0.10       -0.05
## Drama           -1.80     0.28  0.01  -0.02      0.13        0.01
## Family           0.95    -0.29 -0.08   0.18      0.01       -0.10
## Fantasy          0.75    -0.13 -0.02   0.04     -0.05       -0.02
## History          0.35    -0.08  0.06   0.07      0.03       -0.12
## Horror           0.26     0.31  0.02  -0.18     -0.23        0.12
## Music           -0.80     0.36  0.01  -0.16     -0.09        0.13
## Mystery          1.17    -0.01  0.05  -0.10     -0.17        0.04
## Romance         -1.52     0.38  0.01  -0.06      0.02        0.04
## Sci-Fi          -1.48     0.17 -0.07  -0.07      0.09        0.14
## Sport            0.59     0.00  0.09  -0.10     -0.07        0.01
## Thriller         0.03     0.02  0.01  -0.06     -0.02        0.05
## War              0.08    -0.02  0.04   0.00      0.02       -0.04
## Western         -0.59     0.35  0.11  -0.22     -0.09        0.10
```

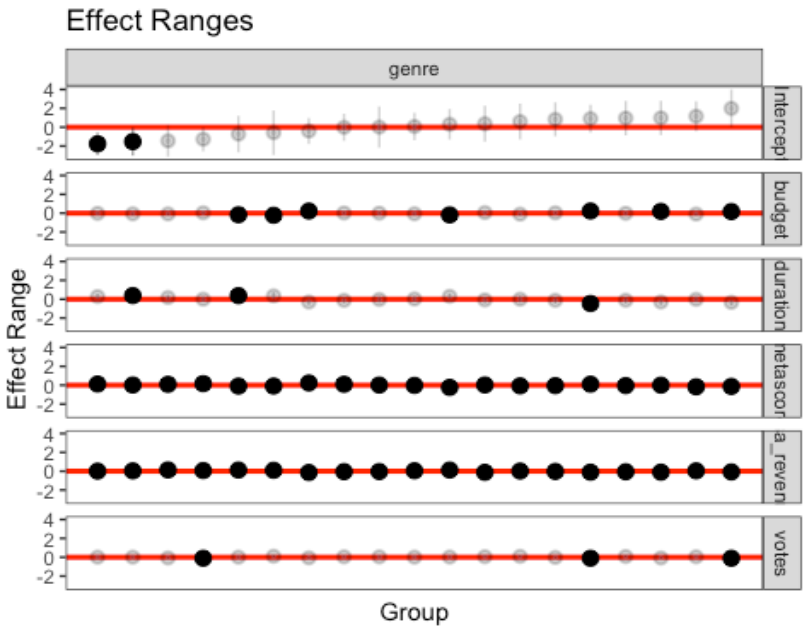

Figure5: Random Effects of Model

The baseline of global revenue is quite different in each genre, which validates that many people prefer to pay for animated movie and adventure movie. Also, ***duration*** and ***budget*** are the two parameters that fluctuate relatively more.

Result

Interpretation

Let's take action movie for example.

```
## $genre
##           (Intercept)    duration     votes     budget    metascore usa_revenue
## Action     -3.0470124  0.12439371 0.2908873 0.49944683  0.378900454   0.4106187
## Adventure  -1.7842812 -0.02917569 0.2466817 0.50165730  0.247494031   0.4431849
## Animation  -0.6818951  0.07847727 0.2455961 0.44536850  0.008821026   0.4704492
## Biography  -1.6959084  0.30454582 0.4330150 0.26163976  0.093440544   0.4880180
## Comedy     -3.9535704  0.42391128 0.2462669 0.31607862  0.303632720   0.6135048
## Crime      -2.6475668  0.29741303 0.3983633 0.29414834  0.234432878   0.5007510
## Drama      -4.4360581  0.71093671 0.3616246 0.25385434  0.263792909   0.5531568
## Family     -1.6833586  0.14028984 0.2719425 0.45808630  0.144067611   0.4443840
## Fantasy    -1.8904194  0.30103177 0.3306478 0.31728587  0.089768999   0.5217783
## History    -2.2865479  0.34771837 0.4112997 0.34462960  0.166888117   0.4269304
## Horror     -2.3741494  0.73956518 0.3732084 0.09296510 -0.093948684   0.6692214
## Music      -3.4364181  0.78499217 0.3635931 0.10897421  0.043984903   0.6731278
## Mystery    -1.4709346  0.41427458 0.3995280 0.17825472 -0.030660140   0.5841482
## Romance    -4.1597977  0.80802145 0.3631872 0.20898944  0.161412826   0.5813341
## Sci-Fi     -4.1171107  0.59487694 0.2798084 0.20066872  0.230563013   0.6845524
## Sport      -2.0487771  0.42858325 0.4441094 0.17170192  0.069444107   0.5608525
## Thriller   -2.6040045  0.44623546 0.3583340 0.21731692  0.115552307   0.5941416
## War        -2.5531773  0.40879543 0.3958107 0.27251307  0.154850266   0.5101802
## Western    -3.2275798  0.77929393 0.4624220 0.05293785  0.048745256   0.6436008
```

I created a formula as follows:

$$log(globalrevenue + 1) = -3.05 + 0.12 \cdot log(duration + 1) + 0.29 \cdot log(votes + 1) + 0.50 \cdot log(budget + 1) + 0.38 \cdot log(metascore + 1) + 0.41 \cdot log(usarevenue + 1)$$

All coefficients are positive, which indicates that all variables have a positive impact on movies' global revenue. For each 1% difference in ***metascore***, the predicted difference in ***global revenue*** is 0.38%. For each 1% difference in ***budget***, the predicted difference in **global revenue** is 0.29%. Both the slope and intercept of variables would vary depending on the genre.

Let me distinguish the different effects of different genres of movies. For animated movies, its base revenue is relatively high, I think it is because the audience group of animated movies is relatively large, especially for children. It is obvious that parents accompany their children to see, or even a family to see. For movies with higher budgets, like ***Action***, ***Adventure***, the relationship between ***budget*** and ***global revenue*** tends to be positively correlated. However, for low-budget movies, ***budget*** have little or no effect on ***global revenue***, and in some cases, they are negatively correlated.

model checking

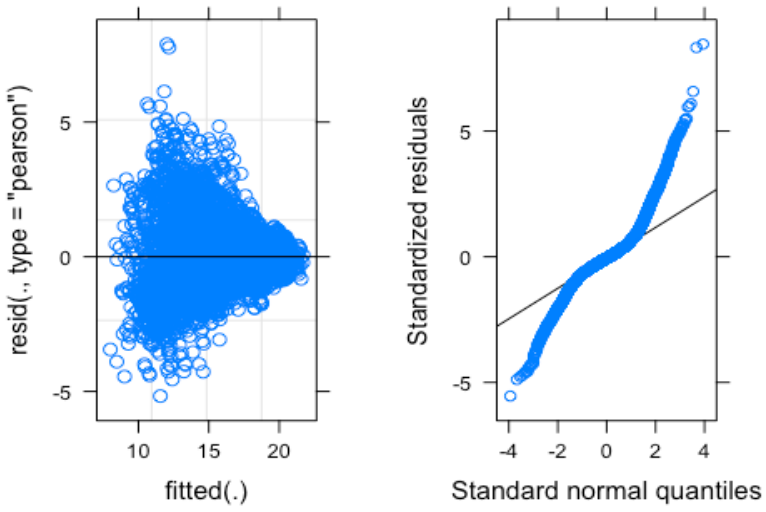

Figure 6: Residual plot and Q-Q plot

From the residual plot, I found that dots are randomly scattered around residuals = 0. Since they are symmetrically distributed and tend to cluster toward the middle of the graph, I conclude that this model is well-fitted for the data. However, by analyzing Q-Q residual plot, the points skew drastically from the line, so I need to consider adjusting my model by adding or removing other variables from the model.

Discussion

In this project, the multilevel model is used to examine the relationship between ***global revenue*** and other numerical variables in different movie genres. In terms of fixed and random effects, all predictors have a positive impact on ***global revenue***. ***usa revenue*** has the largest impact on ***global revenue***, which indicates the outstanding contribution of the U.S. to the world film industry.

In addition, there are still some limitations. For example, the movie schedule, if a movie released on the same day as "The Avengers", the revenue will be much lower. It must be said that competition is an uncontrollable factor. Of course, there are also some market conditions and political factors. None of these factors can be measured by data.

global revenue is also influenced by other non-numerical variables. Firstly, a well-known creative team will undoubtedly bring free publicity to the movie, so I will classify the company by its popularity and add the company's classification variable to the model. Additionally, English movies have a high audience compared to other languages, and I will divide the language category into English and non-English to study the influence of language on ***global revenue***.

Reference

[1]R Bootcamp: Introduction to Multilevel Model and Interactions. https://quantdev.ssri.psu.edu/tutorials/r-bootcamp-introduction-multilevel-model-and-interactions

[2]Explore multilevel models faster with the new merTools R package. https://www.jaredknowles.com/journal/2015/8/12/announcing-mertools

[3]Data Analysis in R. https://bookdown.org/steve_midway/DAR/random-effects.html

Appendix

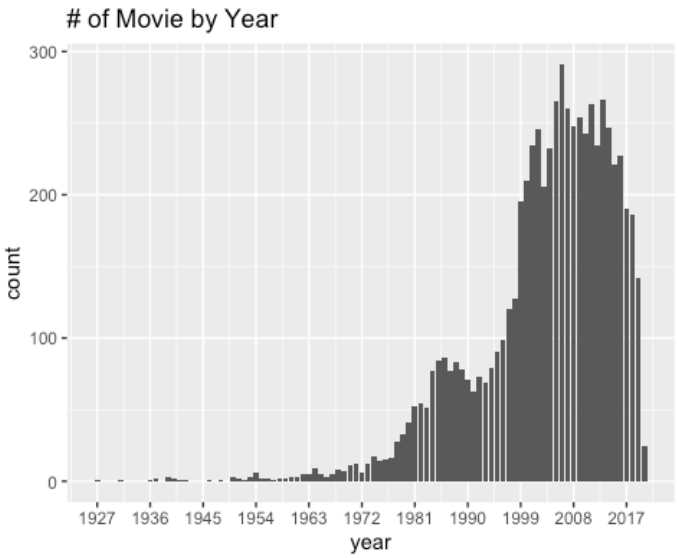

Figure 7: distribution plot

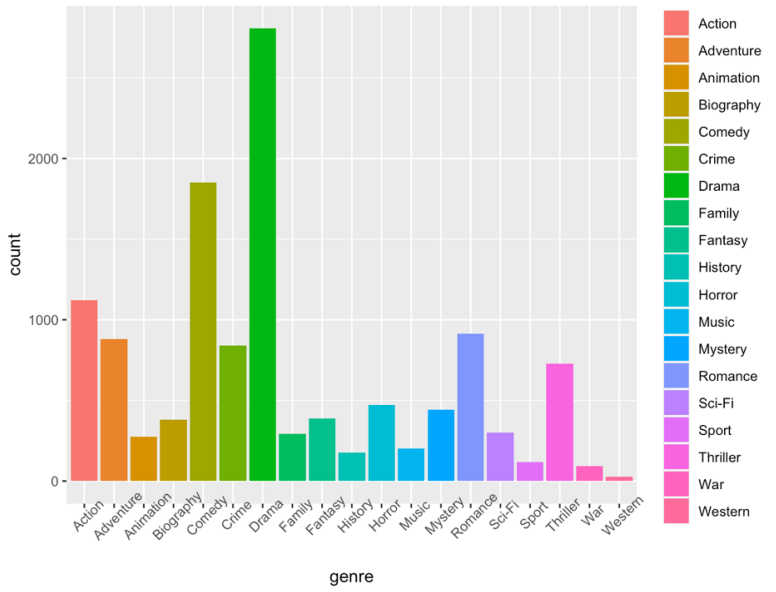

Figure 8: EDA movies type from 1999 to 2019

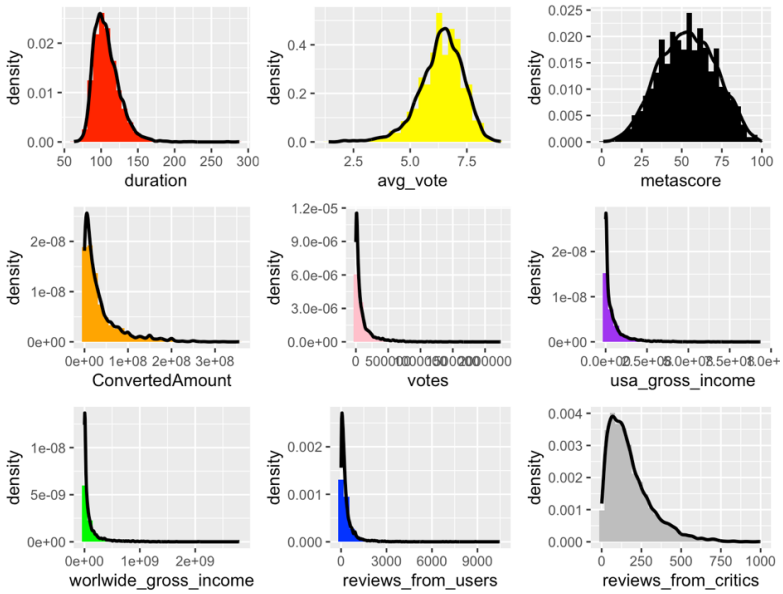

Figure 9: distribution plot

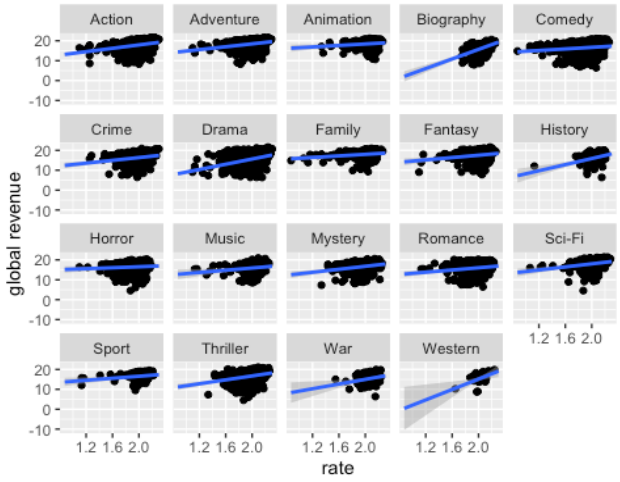

Figure 10: relationship between rate and global revenue in different genre

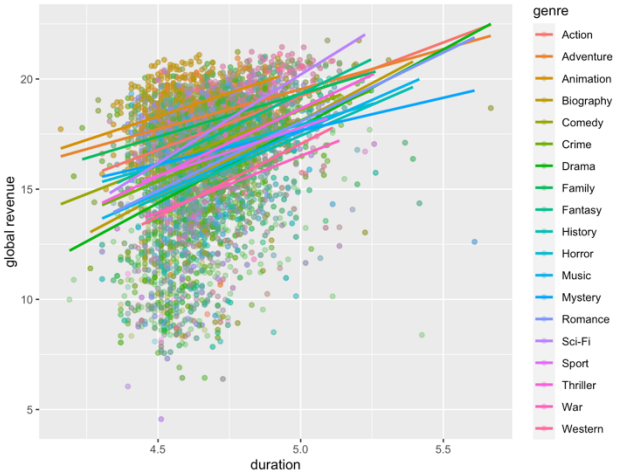

Figure 11: relationship between duration and global revenue

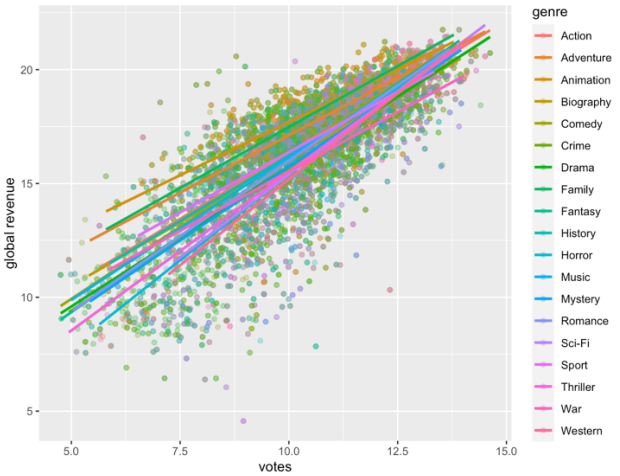

Figure 12: relationship between votes and global revenue

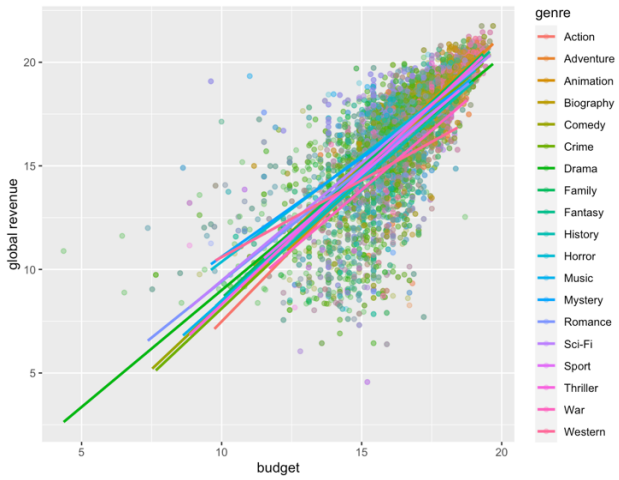

Figure 13: relationship between budget and global revenue

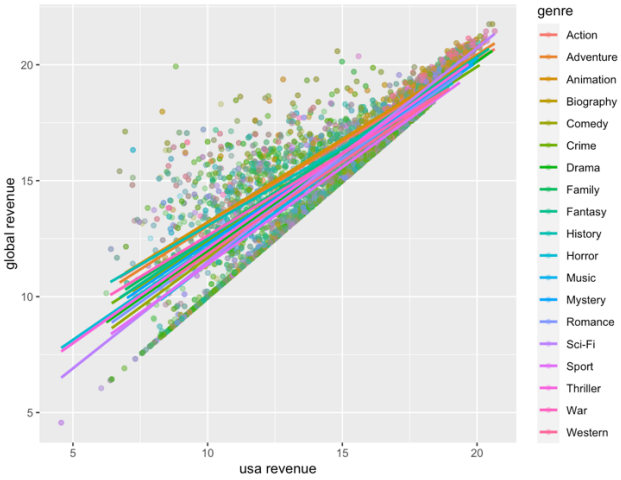

Figure 14: relationship between USA revenue and global revenue